ỨNG DỤNG VÀ CẢI TIẾN HỆ SỐ TƯƠNG ĐỒNG COSINE TRONG XÂY DỰNG VÀ QUẢN LÝ NGÂN HÀNG CÂU HỎI TRẮC NGHIỆM

USING AND IMPROVING COSINE SIMILARITY ALGORITHM FOR BUILDING AND MANAGING QUESTION BANK

Phạm Văn Tính, Nguyễn Thị Phương Trâm
Trường Đại học Nông Lâm TP.HCM, Việt Nam

Ngày toà soạn nhận bài 9/4/2019, ngày phản biện đánh giá 15/4/2019, ngày chấp nhận đăng 03/5/2019.

TÓM TẮT

Ngân hàng câu hỏi trắc nghiệm là thành phần cốt lõi trong hệ thống đánh giá để đảm bảo chất lượng đào tạo trong các tổ chức giáo dục. Các nghiên cứu hiện nay mới chỉ tập trung chủ yếu vào phương pháp tạo ra các đề thi từ ngân hàng câu hỏi có sẵn, mà chưa chú trọng đến việc cần đảm bảo không trùng lặp nội dung các câu hỏi trong ngân hàng câu hỏi. Khi số lượng câu hỏi trong ngân hàng câu hỏi tăng lên thì đồng thời việc quản lý nội dung các câu hỏi cũng trở lên khó khăn. Trùng lặp nội dung trong các câu hỏi là điều khó tránh khỏi. Trong nghiên cứu này chúng tôi ứng dụng hệ số tương đồng Cosine và đề xuất cải tiến giải thuật tính hệ số tương đồng Cosine bằng cách đánh trọng số các từ khóa chính, dùng để phát hiện trùng lặp nội dung câu hỏi trong đề thi hay ngân hàng câu hỏi nhằm đảm bảo các đề thi được phát sinh chính xác hơn.

Từ khóa: *Phát hiện trùng lặp nội dung; Đương đồng văn bản; Hệ số tương đồng Cosine; Hệ số tương đồng Cosine có trọng số; Ngân hàng câu hỏi.*

ABSTRACT

The bank of multiple-choice questions is a core component of the evaluation system to ensure the quality of training in educational institutions. The current research focuses only on the method of creating the exam from the prepared question bank, but it does not focus on the prevention of duplicate material in the question bank. As the number of questions in the question bank increases, the management of questions contents become more difficult and the duplication of question content becomes unavoidable. In this study, we propose using and improving the Cosine similarity algorithm by weighting the keywords (shingles) used to detect the duplicate content of questions in the exams or in question bank to ensure that exams are generated more accurately.

Keywords: *Near Duplicate Detection; Text similarity; Cosine similarity; Weighted Cosine Similarity; Question bank.*

1. GIỚI THIỆU

Lợi thế lớn nhất của thi trắc nghiệm là tính chính xác và chi phí ra đề, chấm thi thấp. Sự nhầm lẫn cũng như khả năng gian lận trong quá trình chấm bài là rất thấp. Đặc biệt với sự trợ giúp của máy tính như hiện nay thì hình thức thi trắc nghiệm càng được áp dụng rộng rãi trong đánh giá môn học.

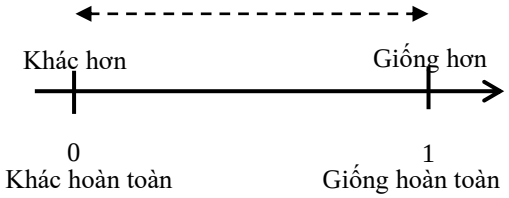
Hiện tại Bộ môn Mạng máy tính và truyền thông có 8/12 môn học sử dụng hình thức thi trắc nghiệm. Trong quá trình tổ chức thi trắc nghiệm chúng tôi ghi nhận được rất nhiều trường hợp có 2 câu hỏi giống nhau trong cùng 1 đề thi.

Liên quan đến thi trắc nghiệm, các nghiên cứu hiện tại chỉ tập trung chủ yếu vào phương pháp tạo ra các đề thi từ ngân hàng câu hỏi có sẵn, mà chưa chú trọng đến việc cần đảm bảo không trùng lặp nội dung các câu hỏi trong ngân hàng câu hỏi [1]-[3]. Trong nghiên cứu này chúng tôi tập trung giải quyết vấn đề trùng

lặp nội dung trong ngân hàng câu hỏi nói chung và các đề thi nói riêng ứng dụng hệ số tương đồng Cosine đồng thời đề xuất cải tiến giải thuật tính hệ số tương đồng Cosine bằng cách đánh trọng số các từ khóa chính.

2. PHÁT HIỆN TRÙNG LẶP NỘI DUNG

Về tổng quan, phát hiện sự trùng lặp giữa 2 tài liệu được xác định thông qua việc tính hệ số tương đồng (similarity) của 2 tài liệu đó. Hệ số tương đồng có giá trị từ 0 đến 1. Giá trị càng tiến đến 1 thì hai tài liệu càng giống nhau và ngược lại giá trị càng gần 0 thì hai tài liệu càng khác nhau.



***Hình 1**. Ý nghĩa của hệ số tương đồng*

2.1 Các bước xác định trùng lặp nội dung

Bước 1: Loại bỏ các từ dừng (stop words) là các từ không chứa thông tin

Bước 2: Tách tài liệu thành các shingle (k-gram hoặc w-gram)

Bước 3: Biểu diễn tài liệu thành tập hợp các shingle duy nhất hay thành vector tần suất. Đây chính là phương pháp vector hóa các văn bản hay nói cách khác biểu diễn tài liệu thành vector.

Bước 4: Tính hệ số tương đồng giữa các tài liệu

Bước 5: Đánh giá hệ số tương đồng để đưa ra kết luận

2.2 Một số khái niệm cơ bản

Stop words: Từ dừng là những từ không chứa thông tin hay có thông tin rất chung chung cần phải loại bỏ trước khi tính toán hệ số tương đồng. Không có danh sách các từ dừng tổng quát. Tùy vào ngôn ngữ mà danh sách các từ dừng này sẽ khác nhau. Trong tiếng Việt stop words có thể là từ đơn (là, mà, v.v) hay cụm từ (đến nỗi, có thể, v.v)

Shingle: Văn bản (tài liệu) là tổ hợp của các ký tự hay các từ. Trật tự của các ký tự hay các từ này cũng có ảnh hưởng đến sự tương đồng của văn bản ví dụ câu "Tôi ăn cơm" khác với "Cơm ăn tôi" mặc dù 2 câu này có các từ giống hệt nhau. Shingling là một phương pháp thể hiện tài liệu thành tập hợp các chuỗi (Shingle) đã bao gồm trật tự của các ký tự trong tài liệu. Nói cách khác Shingle là k-gram trên ký tự hay w-gram trên từ. Ví dụ tài liệu là "I went to work" thì tập hợp 2-shingle trên ký tự là {"I ", " w", "we", "en", "nt", "t ", " t", "to", "o ", " w", "wo", "or", "rk} và tập hợp 2-shingle trên từ (word) là {"I went","went to","to work"}. Thông thường trong phát hiện trùng lặp nội dung sẽ sử dụng w-shingle với hệ số w được lựa chọn từ 2-10.

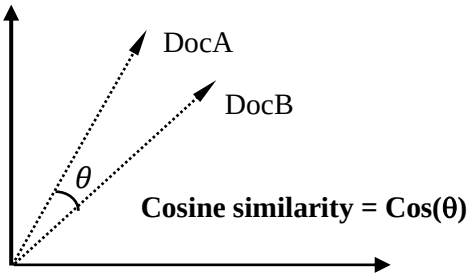
Hệ số tương đồng: là một thành phần cơ bản trong xử lý văn bản. Nó có vai trò quan trọng trong nghiên cứu và ứng dụng liên quan đến văn bản như: phân loại văn bản, tìm kiếm chủ đề, phát hiện và loại bỏ trùng lặp nội dung, tạo và trả lời câu hỏi v.v. Tìm sự giống nhau giữa các từ hay shingle lại là một phần cơ bản trong việc xác định độ tương đồng của văn bản, trên cơ sở đó dùng để xác định sự tương đồng của các câu văn, các đoạn văn hay các tài liệu văn bản. Độ tương đồng của văn bản được ứng dụng trong việc phát hiện sự trùng lặp câu hỏi trong ngân hàng đề thi, phát hiện đạo văn, sao chép nội dung trong bài báo khoa học hay luận văn tốt nghiệp của sinh viên [4]-[6]

Sự tương đồng của các từ có thể xem xét trên hai phương diện: từ vựng (lexical) hoặc ngữ nghĩa (semantic). Các từ tương đồng dạng từ vựng có chuỗi ký tự giống nhau. Các từ tương đồng dạng ngữ nghĩa có chuỗi ký tự khác nhau nhưng có ý nghĩa giống nhau. Ví dụ: "Bắp" và "Ngô" giống nhau về ngữ nghĩa nhưng lại khác xa nhau về từ vựng.

Rất nhiều giải thuật tính hệ số tương đồng trên phương diện từ vựng (Term-Based) được sử dụng như: Jaccard similarity, Euclidean Distance, Dice's Coefficient, Cosine Similarity. Trong số đó Cosine Similarity được sử dụng rộng rãi nhất. Các giải thuật này đều dựa trên việc phân tích chuỗi thành các Shingle sau đó tính độ tương đồng bằng cách so sánh các Shingle thành phần [7].

3. HỆ SỐ TƯƠNG ĐỒNG COSINE

Cosine similarity là một trong những chỉ số phổ biến dùng để xác định tính tương đồng giữa hai đoạn văn bản, được ứng dụng trong tìm kiếm nội dung trùng lặp. Các văn bản được biểu diễn theo mô hình không gian vector.



Hình 2. *Hệ số tương đồng Cosine*

Không gian vector hay số chiều của vector có kích thước bằng tổng số shingle duy nhất trong văn bản. Giá trị mỗi phần tử của vector là tần số xuất hiện của shingle tương ứng trong văn bản. Hệ số tương đồng Cosine là giá trị hàm Cosine của góc giữa hai vector biểu diễn hai văn bản cần so sánh.

Hệ số tương đồng Cosine được tính theo công thức [10]:

$$Cosine\ Similarity = Cos(\theta) = \frac{\vec{A}.\vec{B}}{\left\|\vec{A}\right\|.\left\|\vec{B}\right\|}$$

$$= \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}.\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Trong đó:

- A_i và B_i là các phần tử trong vector A và B của 2 tài liệu DocA và DocB [4],[5],[10]

Để hiểu rõ cách tính, hãy xét ví dụ tính độ tương đồng của 2 tài liệu A, B sau:

DocA: “Ba Một Năm” - “315”

DocB: “Một Hai Ba Một Hai Một Một” - “1231211”

Các bước thực hiện

Bước 1 – Biểu diễn tài liệu thành vector tần suất của các shingle

DocA có 3 shingle phân biệt “1”, “3”, “5”. Mỗi shingle chỉ xuất hiện duy nhất 1 lần

DocB có 3 shingle phân biệt “1”, “2”, “3”. Trong đó shingle “1” xuất hiện 4 lần, “2” xuất hiện 2 lần và “3” xuất hiện 1 lần

VectorA = {(“1”,1), (“3”,1), (“5”,1)}

VectorB = {(“1”,4),(“2”,2),(“3”,1)}

Bước 2 – Chuẩn hóa VectorA, VectorB thành vector có độ dài bằng nhau và bằng độ dài của SetAB là hợp của 2 tập hợp SetA và SetB. Trong đó SetA và SetB là tập hợp các shingle duy nhất của DocA và DocB:

SetAB = SetA ∩ SetB = {“1”, “2”, “3”, “5”} có 4 phần tử duy nhất

VectorA = {(“1”,1), (“2”,0), (“3”,1), (“5”,1)} hoặc đơn giản A= 1 0 1 1

VectorB = {(“1”,4), (“2”,2), (“3”,1), (“5”,0)} hoặc đơn giản B= 4 2 1 0

Bước 3 – Tính hệ số tương đồng Cosine theo công thức (1)

$$\frac{1*4+0*2+1*1+1*0}{\sqrt{\left(1^2+0^2+1^2+1^2\right)*\left(4^2+2^2+1^2+0^2\right)}}$$

$$= \frac{5}{\sqrt{3*21}} = 0.6299$$

Trong trường hợp đánh giá trùng lặp nội dung trên quy mô lớn, cần phải so sánh với số lượng tài liệu lớn, kích thước các tài liệu cũng lớn như trường hợp xác định đạo văn thì kỹ thuật lấy giá trị băm đặc trưng của các tài liệu như SimHash và MinHash [8]-[9] được sử dụng rộng rãi hơn do đơn giản, tốc độ xử lý nhanh và không gian lưu trữ ít - cả một đoạn văn chỉ cần giá trị băm 64 -128 bits. Tuy nhiên phương pháp này có độ chính xác không cao và chỉ phù hợp với tài liệu dài. Với các tài liệu khác nhau hoàn toàn thì SimHash vẫn thường cho giá trị khoảng 0.5 trong khi kết quả mong đợi phải là 0.

Để thấy rõ hơn độ chính xác của SimHash và Cosine xét ví dụ sau:

Trường hợp 1: Hai đoạn văn ngắn và khác nhau hoàn toàn:

V1 = “Hai đoạn văn bất kỳ”

V2 = "Có nội dung khác nhau"

Trường hợp 2: Hai đoạn văn gần giống nhau:

V3 = "Mặt trời mọc ở phía đông"

V4 = "Mặt trời lặn ở phía tây"

***Bảng 1**. So sánh SimHash và Cosine*

	SimHash	Cosine	Chú thích
Trường hợp 1 (V1,V2)	**0.508**	**0.0**	Simhash cho kết quả sai
Trường hợp 2 (V3,V4)	0.766	0.667	

Theo kết quả từ bảng 1, SimHash đã cho kết quả sai trong trường hợp 1. Hai tài liệu nói trên khác nhau hoàn toàn nhưng SimHash vẫn cho kết quả là 0.508 trong khi Cosine cho kết quả đúng là 0.0.

4. CẢI TIẾN PHƯƠNG PHÁP TÍNH HỆ SỐ TƯƠNG ĐỒNG COSINE

Khác với các văn bản thông thường, trong ngân hàng câu hỏi, các câu hỏi thường có nội dung ngắn và đặc biệt nhiều câu hỏi có nội dung tương tự nhau về mặt từ vựng. Trong thực tế đề thi bao gồm nhiều phần. Mỗi phần có các câu hỏi thuộc cùng một chương (phần) và các câu hỏi này thường có nội dung khá giống nhau. Khi ứng dụng hệ số tương đồng Cosine với giá trị ngưỡng từ 0.9-0.95 để kiểm tra nội dung các câu hỏi trong ngân hàng có bị trùng lặp (đã tồn tại) hay không thì cả 8/8 bộ ngân hàng câu hỏi đều phát hiện có sự trùng lặp tuy nhiên khi kiểm tra lại bằng phương pháp thủ công thì không tìm thấy sự trùng lặp.

Hãy xem xét 2 câu hỏi (cặp câu hỏi cùng chủ đề) được trích từ ngân hàng câu hỏi môn học "Mạng máy tính cơ bản" sau:

Câu 1: Trên Internet, phần mềm của người dùng sử dụng **cổng đích nào** để kết nối đến mày chủ **SMTP**?

A) 80

B) 110

C) 25

D) 404

Câu 2: Trên Internet, phần mềm của người dùng sử dụng **cổng đích nào** để kết nối đến mày chủ **POP3**?

A) 80

B) 110

C) 25

D) 404

Với cách tính thông thường 2 câu hỏi này có hệ số tương đồng Cosine = 0.95.

Tương tự, xét 2 câu hỏi khác thuộc ngân hàng câu hỏi môn "Nhập môn hệ điều hành"

Câu 3: Trên **hệ điều hành Linux đĩa cứng** được ký hiệu là had, hdb…và **primary partition** trên đĩa cứng được đánh số là

A) 1

B) 1 đến 4

C) 5 trở lên

D) Tất cả đều sai

Câu 4: Trên **hệ điều hành Linux đĩa cứng** được ký hiệu là had, hdb…và **extended partition** trên đĩa cứng được đánh số là

A) 1

B) 1 đến 4

C) 5 trở lên

D) Tất cả đều sai

Hệ số tương đồng Cosine của 2 câu hỏi này (câu 3 và câu 4) là 0.978. Nếu theo kết quả tính hệ số tương đồng Cosine có thể kết luận 2 cặp câu hỏi trên giống nhau nhưng thực tế đây là các câu hỏi khác nhau hoàn toàn. Trong cặp câu hỏi (1,2) người ra đề đang nói tới cổng kết nối của 2 giao thức khác nhau là **SMTP** và **POP3**, và trong cặp câu hỏi (3,4) người ra đề muốn nói tới phân vùng chính (**primary partition**) và phân vùng mở rộng (**extended partition**) khi chia ổ đĩa cứng.

Mặc dù hệ số tương đồng Cosine đủ tốt và được áp phổ biến trong các ứng dụng khai thác văn bản, nhưng chưa hoàn toàn phù hợp cho bài toán ngân hàng câu hỏi do số từ trong câu hỏi ít và một số câu hỏi có rất nhiều từ giống nhau. Để cải thiện độ chính xác, trong

tài liệu [11] nhóm tác giả đã đề xuất "Khoảng cách tương đồng Cosine có trọng số" (Distance Weighted Cosine Similarity) nhưng thực chất đây là kết hợp 2 phương pháp đo: Hamming Distance và Cosine Similarity, do đó vẫn chưa thể hiện được chủ ý của người ra đề. Vì vậy chúng tôi đề xuất cải tiến giải thuật tính hệ số tương đồng Cosine để phục vụ cho mục đích này.

4.1 Phương pháp đánh trọng số cho hệ số tương đồng Cosine

Với mỗi câu hỏi trong đề thi người biên soạn có thể định nghĩa các từ khóa chính (nếu cần) cùng trọng số tương ứng thể hiện dụng ý của mình. Từ đó biểu diễn thành vector mức độ quan trọng (Vector of Shingle Importance) của các shingle trong tài liệu. Mặc định các shingle có trọng số là 1. Các từ chính sẽ có trọng số >1.

Giả sử vector trọng số của các shingle trong tập hợp các shingle phân biệt của tài liệu A và B là W. Hệ số tương đồng Cosine có trọng số sẽ được tính theo công thức đề xuất sau:

Weighted Cosine (A,B,W) =

$$\frac{\sum_{i=1}^{n} A_i * B_i * W_i^2}{\sqrt{\left(\sum_{i=1}^{n} A_i^2 * W_i^2\right) * \left(\sum_{i=1}^{n} B_i^2 * W_i^2\right)}} \quad (2)$$

Trong đó:

- A_i và B_i là các phần tử thứ *i* trong vector A và B của 2 tài liệu DocA và DocB
- W_i là phần tử thứ *i* trong vector trọng số của tài liệu A

Xét lại ví dụ trình bày trong mục 3:

DocA: "3 1 5"

DocB: "1 2 3 1 2 1 1"

4.1.1 Trường hợp 1- từ khóa nằm trong cả hai tài liệu

Giả sử từ khóa "1" có mức độ quan trọng bằng 5

Bước 1 – Tạo Vector trọng số do người dùng định nghĩa

VectorU = {**("1",5)**}

Bước 2 - Chuẩn hóa Vector trọng số cho tất cả các từ thuộc A ∩ B. Trọng số mặc định cho tất cả các shingle bằng 1:

VectorW = {**("1",5)**, ("2",1), ("3",1), ("5",1)}

Bước 3. Tính hệ sô tương đồng Cosine cải tiến theo công thức (2)

$$\textbf{Weighted Cosine}\ (\mathbf{A},\mathbf{B},\mathbf{W}) = \frac{1*4*5^2+0*2*1^2+1*1*1^2+1*0*1^2}{\sqrt{\left(1^2*5^2+0^2*1^2+1^2*1^2+1^2*1^2\right)*\left(4^2*5^2+2^2*1^2+1^2*1^2+0^2*1^2\right)}}$$

$$= \frac{101}{\sqrt{27*405}} = 0.966$$

4.1.2 Trường hợp 2 - có một tài liệu không chứa từ khóa

Giả sử từ khóa "2" có mức độ quan trọng bằng 5. Từ khóa này không có trong DocA nhưng có trong DocB.

Bước 1 – Tạo Vector trọng số do người dùng định nghĩa

VectorU = {**("2",5)**}

Bước 2 - Chuẩn hóa Vector trọng số cho tất cả các từ thuộc A ∩ B. Trọng số mặc định cho tất cả các shingle bằng 1:

VectorW={("1",1), **("2",5)**, ("3",1), ("5",1)}

Bước 3. Tính hệ sô tương đồng Cosine cải tiến theo công thức (2)

$$\textbf{Weighted Cosine}\ (\mathbf{A},\mathbf{B},\mathbf{W}) = \frac{1*4*1^2+0*2*5^2+1*1*1^2+1*0*1^2}{\sqrt{\left(1^2*1^2+0^2*5^2+1^2*1^2+1^2*1^2\right)*\left(4^2*1^2+2^2*5^2+1^2*1^2+0^2*1^2\right)}}$$

$$= \frac{5}{\sqrt{3*117}} = 0.267$$

Như vậy nếu hai văn bản càng có cùng nhiều từ khóa quan trọng thì càng giống nhau có nghĩa là hệ số tương đồng càng gần bằng 1.0, ngược lại nếu văn bản không chứa từ khóa quan trọng thì càng khác nhau nghĩa là hệ số tương đồng càng gần về 0.0

Bảng 2. *So sánh hệ số tương đồng Cosine và Cosine có trọng số*

Cosine	Weighted Cosine Trường hợp 1	Weighted Cosine Trường hợp 2
0.6299	0.966	0.267

Áp dụng phương pháp tính hệ số Cosine cải tiến cho 2 cặp câu hỏi ví dụ nói trên:

Câu hỏi 1: CH1 = "Trên Internet, phần mềm của người dùng sử dụng cổng đích nào để kết nối đến mày chủ SMTP? . . . "

VectorWCH1 = {("cổng đích",5),
("SMTP",10)}

Câu hỏi 2: CH2 = "Trên Internet, phần mềm của người dùng sử dụng cổng đích nào để kết nối đến mày chủ POP3? . . ."

VectorWCH2 = {("cổng đích",5),
("POP3",10)}

Câu hỏi 3: CH3 = "Trên hệ điều hành Linux đĩa cứng được ký hiệu là had, hdb…và primary partition trên đĩa cứng được đánh số là ... "

VectorWCH3 = {("hệ điều hành Linux", 5), ("đĩa cứng", 5), ("primary partition", 10)}

Câu hỏi 4: CH4 = "Trên hệ điều hành Linux đĩa cứng được ký hiệu là had, hdb…và extended partition trên đĩa cứng được đánh số là ..."

VectorWCH4 = {("hệ điều hành Linux",5), ("đĩa cứng",5), ("extended partition",10)}

Bảng 3. *So sánh hệ số tương đồng Cosine và Cosine cải tiến*

	Không trọng số	Có trọng số	Dụng ý
(CH1, CH2)	0.950	**0.542 (khác hơn)**	**Khác nhau**
(CH3, CH4)	0.977	**0.776 (khác hơn)**	**Khác nhau**

Kết quả bảng trên (bảng 2 và 3) cho thấy phương pháp tính hệ số Cosine cải tiến làm gia tăng sự khác biệt hay tương đồng theo đúng dụng ý của người dùng.

5. ỨNG DỤNG HỆ SỐ TƯƠNG ĐỒNG COSINE VÀ HỆ SỐ TƯƠNG ĐỒNG COSINE CẢI TIẾN TRONG XÂY DỰNG NGÂN HÀNG CÂU HỎI TRẮC NGHIỆM

Hệ số tương đồng Cosine và hệ số tương đồng Cosine cải tiến được ứng dụng để xây dựng phần mềm quản lý ngân hàng câu hỏi thi trắc nghiệm trong 3 chức năng chính:

1. Tạo ngân hàng câu hỏi từ các đề thi trắc nghiệm cũ có sẵn
2. Thêm câu hỏi mới vào ngân hàng câu hỏi
3. Đánh giá, kiểm tra đề thi sau khi phát sinh

Các bước thực hiện:

Bước 1: Dùng hệ số tương đồng cosine với ngưỡng 0.9 để tạo ngân hàng câu hỏi từ các đề thi có sẵn hoặc khi thêm câu hỏi mới vào ngân hàng. Nếu hệ số tương đồng của câu hỏi thêm vào so với tất cả các câu hỏi trong ngân hàng nhỏ hơn ngưỡng (<0.9) câu hỏi mới sẽ tự động được thêm vào ngân hàng.

Bước 2: Trong trường hợp câu hỏi mới có hệ số tương đồng so với tất cả các câu hỏi trong ngân hàng lớn hơn hoặc bằng ngưỡng (>=0.9) sẽ được kiểm tra bằng tay và định nghĩa thêm các từ khóa quan trọng (theo mục 4 – hệ số tương đồng Cosine cải tiến)

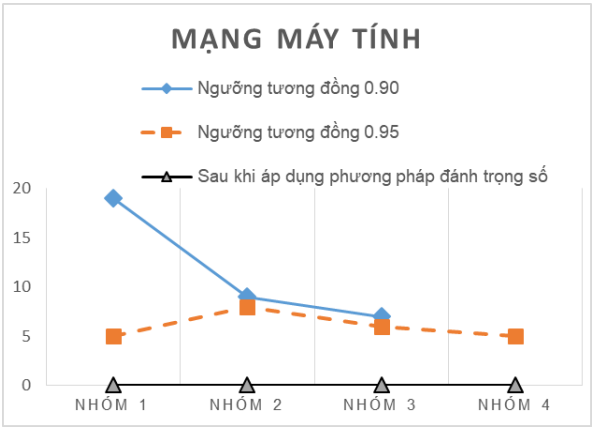
Bước 3: Các đề thi sau khi được phát sinh dùng hệ số tương đồng Cosine cải tiến với ngưỡng 0.95 để kiểm tra đảm bảo không có trùng lặp

Kết quả áp dụng phương pháp trên cho 100 câu hỏi của ngân hàng câu hỏi môn "Mạng máy tính cơ bản" như sau:

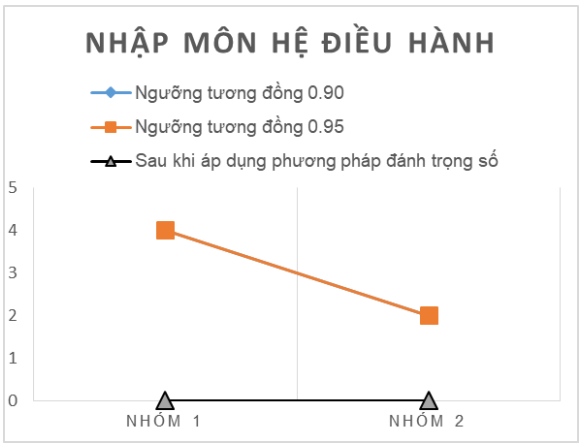
- Bước 1: Với ngưỡng tương đồng 0.90 có 35 câu hỏi tương tự nhau và được chia thành 3 nhóm với số lượng câu hỏi trong mỗi nhóm lần lượt là: 19, 9, 7; Với ngưỡng tương đồng 0.95 chỉ còn 24 câu hỏi tương tự nhau và được chia thành 4 nhóm với số lượng câu hỏi trong mỗi nhóm lần lượt là: 5, 8, 6, 5

- Bước 2 và 3: Sau khi áp dụng phương pháp đánh trọng số cho 24 câu hỏi trong 4 nhóm, số lượng câu hỏi tương đồng là 0.

Áp dụng tương tự với 80 câu hỏi trong ngân hàng câu hỏi môn "Nhập môn hệ điều hành". Với ngưỡng tương đồng 0.90 và 0.95 chỉ có 6 câu hỏi tương tự nhau và được chia thành 2 nhóm với số lượng câu hỏi trong mỗi nhóm lần lượt là: 4, 2. Sau khi áp dụng phương pháp đánh trọng số cho 6 câu hỏi này số lượng câu hỏi tương đồng là 0



Hình 3. *Các nhóm câu hỏi tương đồng môn Mạng máy tính*



Hình 4. *Các nhóm câu hỏi tương đồng môn Nhập môn hệ điều hành*

Bằng phương pháp đánh trọng số các từ quan trọng trong tính hệ số Cosine cải tiến giúp xác định chính xác hơn, phân biệt rõ hơn sự tương đồng hay khác biệt của hai văn bản theo dụng ý của người dùng, đồng thời làm giảm ảnh hưởng của yếu tố từ vựng giúp độ chính xác tiến gần hơn về mặt ngữ nghĩa.

6. KẾT LUẬN VÀ KIẾN NGHỊ

Hệ số tương đồng Cosine được tính dựa vào tần số xuất hiện của các shingle duy nhất trong tài liệu, do vậy sẽ không đánh giá chính xác hai đoạn văn tương tự nhau về ngữ nghĩa nhưng khác nhau về từ vực, hay trường hợp ngược lại hai văn bản rất giống nhau về từ vực (hệ số Cosine lớn) nhưng lại rất khác nhau về dụng ý (ví dụ 2 cặp câu hỏi nói trên). Nói cách khác khi sử dụng hệ số tương đồng Cosine sẽ rất khó thể hiện được đúng dụng ý của tác giả, mà điều này đặc biệt quan trọng trong việc xác định trùng lặp các câu hỏi trong ngân hàng câu hỏi, khi các câu hỏi thuộc cùng nhóm chủ đề rất tương tự nhau, đôi khi chỉ khác nhau một vài từ khóa. Với cải tiến công thức tính hệ số tương đồng Cosine áp dụng phương pháp đánh trọng số từ khóa sử dụng trong xây dựng ngân hàng hỏi thi trắc nghiệm các nhược điểm nói trên đã được khắc phục, đặc biệt đảm bảo việc thêm câu hỏi mới vào trong ngân hàng câu hỏi không bị trùng lặp nội dung.

Theo công thức tính hệ số Cosine cải tiến (2) thì tác động của trọng số đến kết quả phụ thuộc vào cả giá trị của các trọng số cao hay thấp và cả vào độ lớn của tài liệu (số chiều hay kích thước của vector tần suất). Với trường hợp vector tần suất có kích thước nhỏ thì việc tăng giảm giá trị các trọng số có ảnh hưởng lớn tới kết quả. Ngược lại, với vector tần suất rất lớn thì việc tăng giảm giá trị các trọng số lại chỉ có ảnh hưởng nhỏ tới kết quả. Vì vậy tùy vào các trường hợp cụ thể có thể linh hoạt sử dụng phương pháp đánh trọng số bằng giá trị tuyệt đối như các ví dụ ở trên hay tương đối dùng tỷ lệ phần trăm.

Do các câu hỏi trong ngân hàng câu hỏi thường ngắn (kích thước vector tần suất nhỏ) và dễ dàng xác định được các từ khóa quan trọng nên khuyến nghị sử dụng phương pháp đánh trọng số tuyệt đối với giá trị các trọng số giao động từ 3 đến 10. Sau thử nghiệm trên các bộ ngân hàng câu hỏi thuộc bộ môn mạng máy tính và truyền thông xin đề xuất các từ khóa dùng để nhận diện các câu hỏi chung có trọng số là 5, trong đó các từ khóa đặc trưng riêng cho các câu hỏi cùng nhóm là 10. Ngoài ra có thể kết hợp tính độ tương đồng 2 cấp độ sử dụng hệ số Cosine cải tiến có trọng số để xác định không trùng lặp sau đó dùng hệ số Cosine thông thường để tự động phân loại nhóm câu hỏi cùng chủ đề.

TÀI LIỆU THAM KHẢO

[1] Yildirim M., Heuristic optimization methods for generating test from a question bank, Advances in Artificial Intelligence, pp. 1218-1229 (2007).
[2] Yildirim M., A genetic algorithm for generating test from a question bank, Computer Applications in Engineering Education, Vol.18, No. 2, pp. 298 – 305 (2010).
[3] Toan Bui, Tram Nguyen, Bay Vo, Thanh Nguyen, Witold Pedrycz, Václav Snásel: Application of Particle Swarm Optimization to Create Multiple-Choice Tests. J. Inf. Sci. Eng. 34(6): 1405-1423 (2018).
[4] Anand Rajaraman, Jure Leskovec,and Jeffrey D. Ullman,Mining of Massive Datasets, Cambridge University Press, 2014
[5] Felix Naumann, Melanie Herschel, An Introduction to Duplicate Detection, Morgan & Claypool, 2010
[6] Lavanya Pamulaparty, C.V Guru Rao, M. Sreenivasa Rao, A NEAR-DUPLICATE DETECTION ALGORITHM TO FACILITATE DOCUMENT CLUSTERING, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.6, November 2014
[7] Wael H. Gomaa, Aly A. Fahmy, A Survey of Text Similarity Approaches, International Journal of Computer Applications (0975 – 8887) Volume 68 – No.13, April 2013
[8] Anshumali Shrivastava,Ping Li, In Defense of MinHash Over SimHash, Artificial Intelligence and Statistics pp. 886-894 (2014)
[9] Henzinger Monika, Finding near-duplicate web pages: a large-scale evaluation of algorithms, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006
[10] Pratap Dangeti, Statistics for Machine Learning, Packt Publishing, 2017
[11] Li, Baoli: Distance Weighted Cosine Similarity Measure for Text Classification. In IDEAL 2013 proceedings. 10.1007/978-3-642-41278-3_74, 2013

Tác giả chịu trách nhiệm bài viết:
Phạm Văn Tính
Trường Đại học Nông Lâm TP. HCM
Email: pvtinh@hcmuaf.edu.vn